RATIONALIZING THE RATIO DIFFERENCE:

ANALYSIS OF MOLECULAR FACTORS

RELATED TO PRIMATE SKELETAL MUSCLE FIBER TYPE

by

FRANKLIN LEWIS

A THESIS

Presented to the Department of Anthropology
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June 2019

An Abstract of the Thesis of

Franklin Lewis for the degree of Bachelor of Science
in the Department of Anthropology to be taken June 2019

Title: Rationalizing the Ratio Difference: Analysis of Molecular Factors Related to Primate Skeletal Muscle Fiber Type

Approved: _____

Dr. Kirstin Sterner

Bipedalism is a defining human characteristic. Most research on human bipedal evolution has focused on the origins of bipedalism using fossil evidence and bone morphology analysis. Yet few studies have investigated the maximization of bipedal locomotion. In particular, little is known as to how muscular changes influenced human bipedal evolution. In skeletal muscle, slow-twitch fibers produce energy more efficiently and are better suited for endurance activities, whereas fast-twitch fibers consume more energy and are advantageous for activities requiring short bursts of power. When compared to their closest living relatives, the skeletal muscles of bipedal humans have more slow-twitch fibers than those of the quadrupedal apes, but it is still unclear how evolution shaped these patterns. My research addresses this gap in knowledge by characterizing a set of candidate genes that encode proteins that play a role in skeletal muscle physiology and the development of skeletal muscle fiber type. First, I compared the protein-coding sequences of four candidate genes in 25 primates to test if these genes evolved under positive selection. Second, I tested if these genes were differentially expressed in the skeletal muscle tissue of primates with different

locomotor strategies (i.e., quadrupeds and bipeds). The structure of each skeletal muscle fiber is generally conserved between species, whereas the abundance ratio is not. Because genes sequences typically specify the structure of proteins, and the expression of genes specific protein abundance, I predicted that differential gene expression, rather than changes in the coding sequences of genes, is the main source of variation in skeletal muscle fiber-type ratios across species. Preliminary data suggest that these genes are highly conserved across primates and that the expression of these genes in human skeletal muscle tissue is similar to that of other primates. However, I found 2,426 genes that were differentially expressed between human and non-human primates. Reconstructing the evolutionary history of this trait is important for understanding the evolution of human bipedalism and identifying genes involved in skeletal muscle fiber type may also inform our understanding of neuromuscular diseases.

Acknowledgements

So many devoted, passionate people contributed and assisted throughout this journey of research and I admire them all. I want to thank my primary advisor Dr. Kirstin Sterner of the Anthropology department for her constant guidance along the way. She listened to my every question and comment no matter how minute with composure and grace. To state the obvious: this project does not happen without Dr. Sterner. I want to thank both Dr. Barbara Mossberg of the Honors College and Samantha Queeno of the Anthropology department for their support and thorough advice as well. Dr. Mossberg inspires me to become by best academic self, and Samantha's ongoing graduate work was a key point of inspiration for this project. I would also like to thank Dr. Noah Simons, whose bioinformatics tutelage was crucial to the differential expression analysis component of this project. Emily Beck I am also indebted to for her help with the positive selection analyses and providing me access to the Genome cloud server. I also want to acknowledge the work of Dr. Matthew O'Neill, whose study on skeletal muscle fiber ratio inspired this project. To the entire Molecular Anthropology Group for providing support and a healthy working environment, thank you. Finally, I want to the Clark Honors College for providing me with this platform, and all my friends and family for their encouragement of me during my research.

Table of Contents

List of Figures

List of Tables

Introduction

Bipedal Evolution

To characterize any primate — extinct or otherwise — as a hominin, individuals of that species must possess a variety of characteristics, including increased brain capacity and reduced canine size. Above all else, the species must be able to walk upright on two hindlimbs — in other words, it must be bipedal. Since modern humans are the only living hominins today, bipedalism is considered a defining human characteristic. As such, understanding the evolution of bipedality in the hominin lineage is critical to our very identity as human beings. Studies of bone structure in living humans, non-human primates and extinct "hominins" — immediate human ancestors — have identified a number of hominin-specific changes that facilitated more efficient bipedal locomotion. The foramen magnum, the hole in skull that allows the spinal cord to connect to the brain, is more anteriorly placed in humans (i.e., towards the front of skull) than in quadrupeds to support an upright posture (Neaux 2017). Fossil evidence from two extinct hominin species, *Sahelanthropus tchadensis* (Zollikofer et al., 2005) and *Ardipithecus ramidus* (Kimbel et al., 2014), indicate that an anteriorly placed foramen magnum appeared extremely early in hominin evolution, roughly 6 to 7 million years ago. The femur (i.e., thigh bone) became angled inward at the knee and the pelvis became shorter and wider to s3upport the weight of the upper body — instead of distributing weight between hindlimbs and forelimbs as in a quadruped (Tardieu 1994). *Orrorin tugenensis*, another extinct early hominin, was found to exhibit femoral

qualities characteristic of both early hominins and quadrupedal apes, suggesting that changes to the femoral angle was also emerging 7 million years ago (Almecjia 2013).

Although these bipedal features are seen in early hominin species, not everyone agrees that they were bipeds. However most believe species in the genus *Australopithecus* were the first definitive bipeds among primates, (Green et al. 2007, Haile-Selassie et al. 2010, Zipfel et al. 2011) the first species of which appeared approximately 4.4 million years ago (White 1994). The toes of *Australopithecus* and modern humans are shorter and are more arched to support walking and running on the ground, though *Australopithecus* still maintained opposability for climbing in trees (Pontzer 2017). Extinct species from the genus *Homo* including *H. erectus* demonstrate modern human limb properties, including increased surface area of the proximal tibia and distal femur (knee) to support added weight entailing further specialization for bipedalism 1.8 million years ago (Lovejoy 2007, Hatala et al. 2016). Aside from the skull, the skeletal structure of *Homo* populations in Africa and Eurasia 1 million years ago essentially resemble those of living humans today (Pablos et al. 2012, Sawyer et al. 2005, Arsuaga et al. 2015).

Using the trail of evidence described above, several hypotheses exist to explain the origin of human bipedalism (Ko 2015). The "savannah model" suggests that a climate shift during the time of early hominins turned what was normally a heavily forested environment into a grassland savannah (Rodman 1980). This hypothesis suggests that the upright posture associated with bipedality gave savannah-dwelling hominins a fitness advantage, allowing them to detect, and evade, predators lurking in the grass. Extending from the savannah model, the "thermoregulatory model" argues

that a bipedal posture helped early hominins regulate their internal body temperature given the decreased shade cover on grasslands (Wheeler 1992). Bipedal posture decreases an individual's exposure to solar radiation, as well as increasing convective and evaporative heat loss (Wheeler 1991).

Other hypothesizes of human bipedal origins center around feeding and food acquisition. The "postural feeding model" claims that bipedal posture helped early hominins acquire food from low hanging or thin branches – which might explain why the lower limbs of early hominins were structured for bipedal locomotion, but their upper limbs are characteristic of arboreal graspers (Hunt 1994). Hominins that could reach more food by standing upright would have higher evolutionary fitness than individuals limited to quadrupedal posture. The "provisioning model" incorporates fossil evidence of sexual dimorphism between early hominins to explain the evolution of hominin bipedality. According to this model, bipedality gave a fitness advantage as it freed up the hands of male hominins to carry food back to their female partner, increasing their chances of siring offspring (Lovejoy 1988).

However, recent findings have cast doubts on all of these theories, and the question is still debated (Reynolds et al. 2015). Paleontological and biogeochemical studies suggest that our earliest hominin ancestors inhabited forested environments where they could have spent a certain amount of time in the trees (Senut 2017), which casts doubt on the savannah model and related theories. For example, the provisioning model assumes that hominins were adjusting to a lack of food in a savannah environment, which prompted provisioning behavior (Lovejoy 1988). If hominins were still living in forested environments, then perhaps provisioning behavior did not

manifest. Additionally, another more recent model that is still contested is the "wading model," which hypothesized that early hominins began wading into rivers in search of food (Kuliukas 2002, Niemitz, 2002). This model uses evidence of femur abduction and dentition microwear in *Australopithecus* that was previously unexplained by the other models.

Skeletal Muscle Fiber

While most research in this area has concentrated on disentangling the origins of hominin bipedality and focused on skeletal morphology, less research has examined the role of soft tissue (i.e., skeletal muscle) changes in maximizing the efficiency of bipedal locomotion. Humans' running endurance exceeds those of other primates, which suggests that selection for locomotive endurance traits — particularly long-distance running — played a key role in shaping the human lineage (Carrier 1984). Fossilized animal bones beginning from 2.5 million years ago show evidence of stone tool butchery markings, perhaps indicating a dietary shift to meat that would have likely required Homo species to occupy larger geographic ranges (Carbone et al. 2005, Pontzer 2012). If humans were hunting over larger ranges, locomotive efficiency likely increased (Bramble 2004, Liebermann 2014). Besides the skull, early human populations in Africa and Eurasia 1 million years ago so similar to modern humans (Pablos et al. 2012, Sawyer et al. 2005, Arsuaga et al. 2015) that economy and endurance were most likely equivalent to living humans. One source of human's superior endurance is a greater amount of slow-twitch fibers in skeletal muscle (O'Neill 2017). Additionally, human muscle fibers are shorter than average for primates, which also indicates a preference for repetitive, economic movement (O'Neill 2017).

Each skeletal muscle is composed of many thousands of fibers. These fibers are responsible for creating energy to move actin filaments that when synchronized produce muscle contractions required to move (Frontera 2015). In primates, those fibers generally come in two types: slow-twitch fibers (also referred to in other literature as type I) and fast-twitch fibers (also referred to in other literature as type II) (Talbot 2016). Slow-twitch fibers generate energy efficiently and are fatigue resistant, whereas fast-twitch fibers generate large amounts of energy rapidly but cannot sustain energy production over time (Peter 1972, Talbot 2016). The characteristics of a marathon runner and a weightlifter is a helpful metaphor for understanding the difference in the two fiber types: marathon runners rely more on slow-twitch fibers during a long race, whereas weightlifters rely more on their fast-twitch fibers to move a heavy object for a short period of time.

Humans are born with a higher percentage of slow-twitch fibers than fast-twitch fibers in their skeletal muscles, particularly in the hindlimb (legs) (Umberger 2003). While humans can alter their physique through specific forms of exercise, the ratio of slow-twitch fibers to fast-twitch fibers remains the same. Those phenotypic changes from exercise are attributed to converting a third type of hybrid muscle fiber to either fast- or slow-twitch (Klitgaard 1990). Many of humankind's closest primate relatives have more fast-twitch fibers than slow-twitch fibers in contrast to humans (O'Neill 2017). While fiber ratios can differ between individual muscles, O'Neill and his colleagues found that humans overall have close to 60% slow-twitch fibers in their skeletal muscle, whereas chimpanzees have around 30% slow-twitch fibers. They found similar results in the existing primate skeletal muscle fiber ratio literature — the only

non-human primate with greater slow-twitch fiber abundance was the slow loris (O'Neill 2017), who's slow climbing locomotion strategy (Nekaris 2001), would seem to lend itself to favor slow-twitch fibers. Additionally, the magnitude of chimpanzee-human differences in skeletal muscle fiber type ratio is greater than any induced change through intense athletic training (O'Neill 2017). This difference in muscle types is consistent with behavioral data between humans and other primates. O'Neill's findings prompted two questions: 1) When during hominin evolution did this pattern emerge? and 2) Were changes in fiber type ratios selected for by natural selection during hominin evolution?

Genetics of Fiber Type/My contribution

In order to investigate the evolutionary history of fiber type in hominins it is important to first identify what regions of the "genome" — the programming code for nearly all physical characteristics an individual — control this trait. The stark difference in skeletal muscle fiber ratios between humans and other primates is likely caused by genetic or regulatory differences in humans. However, pinning down the genetic factors associated with each skeletal muscle fiber type is difficult because of the complex nature and structure of skeletal muscle fibers, and the inherent challenges in connecting genotypes to complex phenotypes. Most research suggests that several genes are likely contributing to the overall muscle fiber ratio (Scheffanio 2011, Spangenberg 2003), which is typical of most human traits. My research will address this gap in knowledge by helping to characterize the molecular underpinnings of skeletal muscle fiber type. Specifically, I investigated two explanations as to why slow-twitch fibers are more abundant in human skeletal muscle than in other primates: genetic variation and

differential gene expression. If the DNA sequence of genes related to skeletal muscle fiber type are different in humans relative to other primates, then genetic variation may explain the unique human fiber abundance ratios. Alternatively, if the expression — the rate of transcription of the genes from DNA into RNA (i.e., how and when genes are activated and used in the body) — of fiber type genes is different in humans, then differential expression may explain human's fiber type ratio.

Although research into genetics of skeletal muscle fiber type is still lacking, studies with mice have produced several genes of note. In mice, the genes *Smnl1*, *Aox1*, *Dci* and *Casq2* are expressed primarily in slow-twitch muscle fibers, while the genes *Myoz1*, *C2cd21*, *Srebf1* and *Gapdh* are expressed primarily in fast-twitch fibers (Chemello 2011). In addition, the expression of genes *Ckm*, *Pparg*, and *Fnip1* is correlated with an increase in slow-twitch skeletal muscle fibers in mice (Naya 2000, Johnson 1989, Wang 2004, Reyes 2015). I selected the human orthologs of 4 of these candidate genes (*PPARG*, *FNIP1*, *AOX1*, and *MYOZ1*) for my study. *PPARG* codes for a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. PPARs form heterodimers with retinoid X receptors (RXRs) and regulate transcription of various genes. Additionally, *PPARG* has been implicated in the pathology of numerous diseases including obesity, diabetes, atherosclerosis and cancer (Li 2017). *FNIP1* encodes a protein that binds to folliculin — a tumor suppressor protein — and to AMP-activated protein kinase (AMPK) in humans. The encoded protein participates in the regulation of cellular metabolism and nutrient sensing by modulating the AMPK and target of rapamycin signaling pathways (Baba, 2006). *AOX1* encodes aldehyde oxidase, which produces hydrogen peroxide and, under certain

conditions, can catalyze the formation of superoxide. Aldehyde oxidase is a candidate gene for amyotrophic lateral sclerosis (ALS or Lou Gehrig's disease) (Garattini 2008). The protein encoded by *MYOZ1* is primarily expressed in human skeletal muscle and belongs to the myozenin family. Members of this family function as calcineurin-interacting proteins that help tether calcineurin to the sarcomere of cardiac and skeletal muscle. They play an important role in modulation of calcineurin signaling (Lin 2014). Given the slow-twitch fiber specificity of *PPARG*, *FNIP1*, and *AOX1* in mice, I expect to see greater expression of these genes in human skeletal muscle tissue compared to non-human primate muscle tissue. Since *MYOZ1* is expressed in mice fast-twitch fiber, I expect to observe greater expression of these genes in non-human primate skeletal muscle tissue compared to human muscle tissue.

Research Objectives and Questions

QUESTION 1: Is the base pair construction of the candidate sequences significantly different in humans and are the sequences under positive selection?

Objective 1a: Compare the protein-coding sequences of four candidate genes between humans and other primates to test if significant variation exists in the human sequences.

$H1_{a1}$: The protein-coding sequences of *AOX1*, *FNIP1, MYOZ1* and *PPARG* will be conserved in non-human primates but derived (i.e., different) in humans.

$H1_{a2}$: The protein-coding sequences of *AOX1*, *FNIP1*, *MYOZ1* and *PPARG* will be highly variable across the primate phylogeny.

$H1_{a3}$: The protein-coding sequences of *AOX1*, *FNIP1*, *MYOZ1*, and *PPARG* will be highly conserved across the primate phylogeny.

Objective 1b: Test if candidate genes are under positive selection.

$H1_{b1}$: There will be evidence of positive selection in, *AOX1, FNIP1, MYOZ1*, and *PPARG* in primates.

$H1_{b2}$: There will be no evidence of positive selection acting on, *AOX1 FNIP1*, *MYOZ1* and *PPARG* in primates.

QUESTION 2: Does expression of selected genes in skeletal muscle tissue vary between humans and non-human primates?

Objective 2a: Test if the four candidate genes are expressed differently in the skeletal muscle tissue of quadrupedal (non-human) primates vs. humans.

$H2_{a1}$: The expression of *AOX1*, *FNIP1*, and *PPARG* will be higher in humans relative to other primates.

$H2_{a2}$: The expression of *MYOZ1* will be lower in humans relative to other primates.

$H2_{a3}$: The expression of these genes in human muscle will fall within the range of the expression of these genes in non-human primate muscles.

Objective 2b: Determine which biological functions are over-represented by differentially expressed genes in human skeletal muscle tissue.

$H2_{b1}$: Genes related to skeletal muscle fiber will be differentially expressed between humans and non-human primates.

Methods

Part I: Testing for Genetic Variation

Sequence Collection

In order to analyze the protein-coding regions of my four candidate genes, I first downloaded orthologous sequences for all primates currently available in public genomic databases. These primate sequences correspond with the candidate human genes. The majority of sequences came from Ensembl, an online genome browser containing vertebrate genomes (Zerbino et al 2018). I searched each candidate gene by name and then downloaded all known one-to-one orthologues for that gene present in publicly available primate genomes. The primate genomes available at the time of download included: *Homo sapiens* (humans), *Pan troglodytes* (common chimpanzee), *Pan paniscus* (bonobo), *Gorilla gorilla* (western gorilla), *Pongo abelii* (Sumatran orangutan), *Nomascus leucogenys* (northern white-cheeked gibbon), *Papio anubis* (olive baboon), *Cercocebus atys* (sooty mangabey), *Chlorocebus sabaeus* (green monkey), *Rhinopithecus bieti* (black snub-nosed monkey), *Rhinopithecus roxellana* (golden snub-nosed monkey), *Colobus angolensis palliatus* (Angola colobus), *Macaca fascicularis* (crab-eating macaque), *Macaca mulatta* (rhesus macaque), *Macaca nemestrina* (pig-tailed macaque), *Mandrillus leucophaeus* (drill), *Saimiri boliviensis* (Bolivian squirrel monkey) *Cebinae* (capuchin monkey), *Callithrix jacchus* (marmoset) *Otolemur garnettii* (bushbaby), *Aotus nancymaae* (Ma's night monkey), *Carlito syrichta* (Philippine tarsier), *Microcebus* (mouse lemur) (see Appendix A for specific gene IDs used). I also included *Rattus rattus* (rat) and *Mus musculus* (house mouse) as outgroup

species. Sequences were downloaded as a multi-species FASTA file, one per gene, of unaligned, coding sequences. I also searched the National Center for Biotechnology Information (NCBI) database for additional primate sequences using NCBI blast (Altshul 1990 et al.) for each candidate gene. I downloaded two additional species' sequences, *Theropithecus gelada* (gelada baboon) and *Piliocolobus tephrosceles* (Ugandan red colobus). I added these sequences to the multi-species FASTA file because they were not yet available on Ensembl, although they are available now.

Sequence Alignments

To align the sequences, I uploaded each gene's FASTA file to an online alignment program called Clustal Omega (Chojnacki et al 2017). Clustal Omega is a free online program run by the European Bioinformatics Institute to generate alignments between three or more genetic sequences. This program produced a DNA alignment for each gene and output these alignments in NEXUS format. To visualize and work with the NEXUS files, I used Mesquite, a free software program available for download that helps organize biological data in a variety of ways (Maddison 2018). For my purposes, I used Mesquite to view and edit Clustal Omega's alignment. Several things might need to be realigned, including gaps in the sequences that do not conform to proper codon boundaries. My general procedure in Mesquite is as such: I first set the codon index to read the sequences in groups of three, with the first base being position 1, the second position 2, and the third base position 3. On the fourth base, the index resets, so the fourth base is read as the beginning of a new codon, and so on and so forth. I then set the program to color each codon based on its corresponding amino acid. This allowed me to immediately recognize when a sequence was out-of-alignment with

the others, because the matching columns of corresponding color would be interrupted by an out-of-place color, continuing on for several columns. Scanning for premature stop codons, massive frameshifts or gaps are key indicators a given sequence is misaligned. The most common challenge I encountered that produced misalignment was insertions and deletions of whole codons in a sequence. Insertions or deletions of whole codons do not change what the downstream codons code for, but they do often trick Clustal Omega into incorrectly aligning sequences. To deal with codon insertions and deletions, I would use the break tool in Mesquite to indicate a break in a sequence, then manually adjust the sequences to enforce proper codon boundaries. This preserves the alignment upstream of the codon insertion or deletion and fixes the alignment downstream of the inserted or deleted codon. With the alignments completed, I was able to use these data to run two different types of evolutionary analyses: phylogenetic analyses and tests of positive selection.

Phylogenetic Analyses

In order to determine if the genetic sequences of these candidate genes varied significantly in humans relative to other primates, I constructed gene trees using the alignments obtained from Mesquite. A gene tree is a way to visualize the relationships of the sequences present in your dataset. Gene trees are constructed through analyzing sequence variation in a gene or multiple genes and using this information to infer the evolutionary history of the locus. Genes that are highly conserved have little phylogenetic signal (i.e., variation) and will typically produce trees that are unresolved with a "comb-like" branching pattern: few points of common ancestry, or nodes and with low statistical confidence in the nodes that are present. Genes that are highly

variable will typically produce a tree that is able to resolve the relationships between the sequences in the dataset. High-confidence trees with many nodes are produced by such genes.

To construct my gene trees, I used MEGA7 software to run a likelihood analysis. MEGA is a free program available for download with the ability to conduct a wide array of phylogenetic manipulations and tests (Kumar 2016). I used MEGA7 to produce gene trees with bootstrap values for each node of the tree. For a given gene, I first uploaded the Mesquite NEXUS file. Once the file has been aligned to MEGA7 stipulations, I employed the "find best DNA/protein model" function in MEGA7. This tested my data against 28 evolutionary models to see which model best described the variation between each species' version of a gene. The model with the lowest Bayesian Information Criterion score was selected as it was determined that to best represent the substitution model. I then used this model to infer a "maximum likelihood tree," from the dataset. The maximum likelihood tree I labeled with bootstrap values, which are calculated by MEGA7. Bootstrap values represent the confidence in a node, or in other words, the amount of times a given node was predicted over 100 tests. The final result is a phylogenetic tree that I can use to assert the impact these genes' sequences had on primate skeletal muscle evolution.

Tests for Positive Selection

The final step to part I of my analysis was to test if the candidate genes are under positive selection. Positive selection refers to natural selection favoring (or "selecting," hence the name) advantageous genetic variants in a population. To test for positive selection, I used software called Phylogenetic Analysis by Maximum

Likelihood (PAML) (Yang 2007) which contains several evolution analysis programs. The program I used, codeml, compares how different evolutionary models best fit the observed variation in a given gene. Codeml requires three pieces of input: sequence alignment file, a species tree, and a control file. The sequence alignment file was generated for Objective 1a (see above). A species tree represents known species relationships, unlike a gene tree which represents the relationships of the DNA sequences themselves and is inferred directly from the data. I used previously published data to compile the species tree (Sterner *pers comm*). The control file sets the parameters for the analyses, including which evolutionary models are being compared, the assumed random mutation rate, and other statistical stipulations. The control file also tells the program which alignment and tree files to run and where to find the files. Once I assembled the alignment file, the species tree file and the control file, I placed all three files into the same folder as the codeml file. This folder was located on Genome, a cloud computing service controlled by the Cresko Laboratory at UO. I used a free application called FileZilla to organize and access these files. I then used the Terminal application on my computer to log into the Genome computer, changed my directory to my "PAML" folder, created a screen to run my analyses in the background and executed the code "./codeml" to commence the analysis. I repeated these steps for each gene so that after setting up four screens for each gene, I obtained one output file for each gene. The primary result I was interested in from the output files was the log-likelihood scores (lnL) for both the neutral model (M7) and the positive selection model (M8). I used the Log-Likelihood Ratio Test [LRT=2(lnLp-lnLn)] to test if the M8 model was a significantly better fit to my data than the M7 model. The LRT values for

each gene can be used in a chi-squared test to calculate the statistical significance of the result based on the degrees of freedom (df). Df is calculated by subtracting the number of parameters of the positive selection model by the parameters of the neutral model. The df for my analyses was 2. I then calculated the p-value for each of my tests.

Part II: Testing for Differential Expression

Transcriptome Data Collection

Gene expression is measured through counting RNA transcripts of a given gene that are present in a cell or tissue sample. The RNA sequence counts for human and primate tissue samples (including skeletal muscle) is compiled in an online public database, the Non-human Primate Reference Transcriptome Resource (NHPRTR) (Peng et al. 2014). For my analysis, I used data from "set II" which were reads from specific tissues of eleven different primates: human, chimpanzee, rhesus macaque, Japanese macaque, crab-eating macaque (Chinese and Mauritian variants), pig-tailed macaque, olive baboon, sooty mangabey, common marmoset, squirrel monkey and mouse lemur. The data file I used was originally prepared by Noah Simons and is publicly available on GitHub for download. Simons extracted a subset of the total data exclusive to skeletal muscle fiber.

Metadata Creation and Loading File into RStudio Project

Besides the raw counts data file, I created a file to group the species based on the comparisons of gene expression I wanted to make. Referred to as the metadata file, I made a plain text file that used a simple yes/no system for whether a species was human or non-human (see Appendix B for a detailed version of the metadata file). DESeq2

uses this metadata file to group the species into human (Y) or non-human (N) groups, then determines which genes in each group are differentially expressed relative to the other group. This kind of "one-to-all" comparison potentially artificially inflates how different the expression of the single species is relative to the others, but given the limitations of the data available to me, I opted for this technique instead of a one-to-one comparison (e.g., human versus chimpanzee). Next, I loaded my raw data and metadata files in to a new project in RStudio. I then loaded DESeq2, ggplot2, RColorBrewer, pheatmap and DEGreport, which contain all the pre-written programs necessary for my analysis.

Normalization and Quality Control

Before running the differential expression analysis, I first needed to normalize the expression counts data set. In this context, normalization refers to scaling the raw count values to account for factors outside of my control, primarily sequencing depth and RNA composition. Sequencing depth refers to the amount of times the RNA of a gene was counted during sequencing — the more times a gene was counted, the higher the observed genetic expression, regardless of the actual genetic expression. I also had to account for RNA composition differences between the samples. To clarify, DESeq2 does not use normalized counts as input, rather it uses the raw counts and models the normalization inside the Generalized Linear Model (GLM). However, DESeq2 does not have visualization options, so I still needed to normalize the data for ggplot2 and pheatmap, which I used to produce my plots. My first step was to execute code that checked if the species names in the raw counts file matched those in the metadata file. I next created a DESeqDataSet object using the counts file, the metadata file, and a

design formula. The design formula tells DESeq2 which column or columns in the metadata table to focus on and how it/they should be used in the analysis. I executed code that estimated size factors for my samples, then executed code that applied those size factors to the appropriate counts to create normalized counts. I then saved the normalization output in a separate file.

Differential Expression Analysis and Visualization

DESeq2 has two main functions. It first normalizes the RNA count data to correct for differences in library sizes or RNA composition between samples. With normalized data, DESeq2 can determine differential expression between genes and across species. For my purposes, I needed only two lines of code to initiate DESeq2. I needed to execute code that created a DESeq2 object that specifies the location of my raw counts, metadata, and a design formula. With the object created, I specified DESeq2 to run on that object. DESeq2 calculated differential expression using log fold change, which is a statistical method that measures how much a quantity changes from a starting value to a final value.

DESeq2 produces a list of genes that are differentially expressed in humans relative to the other ten primate species. After obtaining the list, I applied a fold change threshold to the list to try to eliminate any extreme outliers. I created plots of expression for my four candidate genes, and I created a plot of all differentially expressed genes using R functions plotMA, plotCounts and ggplot2. I also created a heat map of differential expression across species using the R function pheatmap. Finally, I saved my results in three files: a list of up regulated genes, a list of down regulated genes, and the complete list of differentially expressed genes.

Enrichment Analysis

The final piece to part II of my analysis was examining the functions of the genes that were differentially expressed, using an enrichment analysis. Because the list of differentially expressed genes is so large, it is helpful to group the genes into functional categories and observe which categories are overrepresented in my dataset rather than going gene-by-gene. To do this enrichment analysis, I used the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang 2009). DAVID has a wide variety of functions, but I used it to conduct a gene ontology analysis to test for biological processes, cellular components or molecular functions that were overrepresented in my up-regulated and down-regulated differentially expressed gene list.

Results

Candidate gene sequences are conserved between humans and other primates

In general, visual observation of the aligned sequences and construction of gene trees revealed conservation of candidate gene sequences between humans and other primates, rejecting hypotheses H1A1 and H1A2 and supporting hypothesis H1A3. All trees predicted known evolutionary relationships between primates. Most important to my hypothesis, humans were not separated out from other primates, let alone the rest of the apes. This indicates that the human candidate gene sequences did not vary significantly from other primates. Tree branch lengths of most primate species were short as well — further suggesting a lack of variation between the sequences (Figure 1). The tree with the highest confidence bootstrap values overall was *AOX1* (Fig 1A). Specific nodes of the other trees had high bootstrap values (greater than 0.85), but overall did not have as much confidence nor did the other trees resemble known relationships as well as *AOX1* (Fig 1B-D).

Figure 1: Gene Trees for Candidate Genes

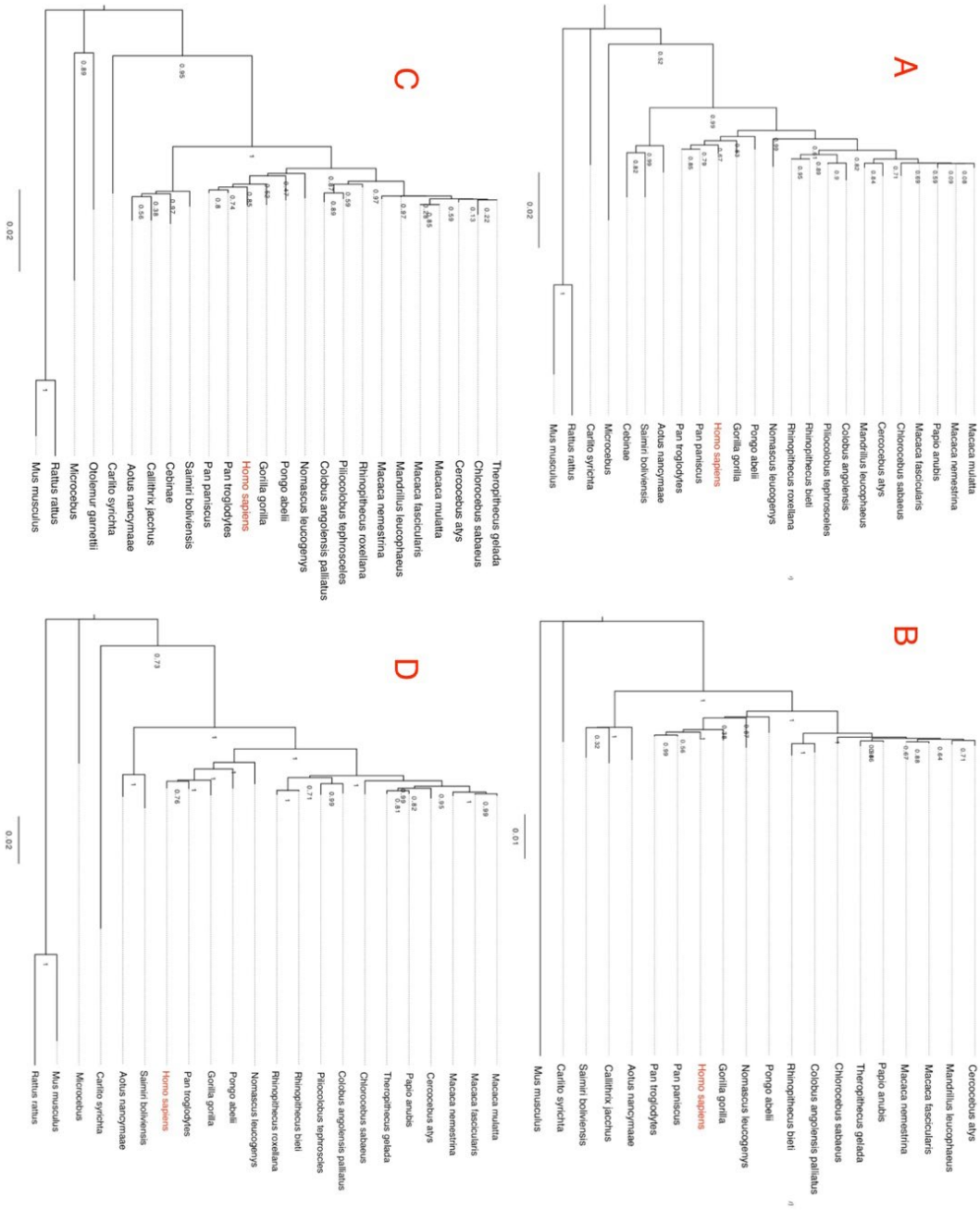


Figure 1: MEGA7-constructed gene trees for AOX1 (A), PPARG (B), MYOZ1 (C) and FNIP1 (D) displaying bootstrap values for each node. Branch lengths indicate projected evolutionary time, meaning the amount of time between a species and its last common ancestor with an adjacent species. Bootstrap values indicate the confidence of the relationship, i.e.: how likely is that relationship true (greater than 0.80 indicates high confidence)

AOX1 and FNIP1 are under positive selection

PAML produced site model analysis results for both a neutral model of selection (M7) and a model of positive selection (M8) (Table 1). I was able to rule out the null hypothesis (neutral selection model) for both *AOX1* ($p < 0.001$) and *FNIP1* ($p = 0.02$), indicating positive selection is acting on these two genes. For *MYOZ1* and *PPARG*, I could not rule out neutral selection ($p > 0.95$), which suggests that positive selection is not acting on these genes. This partially supports my hypothesis H1B1. Somewhat surprisingly, Bayes Empirical Bayes (BEB) analysis (Yang, Wong & Nielsen 2005) was unable to identify specific sites under positive selection in *AOX1*. In *FNIP1*, BEB analysis detected one amino acid site under positive selection: position 850.

Table 1: Positive Selection Analysis Values

Gene	Model	lnL	$2(\ln L_8 - \ln L_7)$	df	p-value
AOX1	M7	-14664.05	18.14	2	**> 0.0001****
	M8	-14654.98			
FNIP1	M7	-8036.82	7.48	2	**0.02***
	M8	-8033.08			
MYOZ1	M7	-2558.50	> 0.0001	2	< 0.99
	M8	-2558.50			
PPARG	M7	-3967.63	0.0950	2	0.9535
	M8	-3967.59			

Table 1: Likelihood ratio test values as calculated from site model analyses and significance estimates based on chi-squared test. ** = highly significant, * = significant

Part II

Human skeletal muscle tissue transcriptome differs from other primates

After normalizing the raw counts data, I was able to make some initial inferences about human skeletal muscle gene expression relative to other primates. Using principal component analysis (PCA), I found that the human transcriptome in skeletal muscle tissue varied significantly from that in other primates (Fig 2). Focusing on the PC1 scale — which accounted for 48% of the variation between species — the separation between humans and the next closest non-human primate (chimpanzees) was approximately 6 times that of the distance between chimpanzees and the rest of the non-human primates based on spatial position on the plot. The rest of the non-human primate skeletal muscle tissue transcriptomes were nearly equivalent in transcriptional space according to PC1. Focusing on PC2 — which accounted for 15% of the observed variation — chimpanzees appear to vary most from the rest of primates, but the level of separation is small compared to human's separation in PC1. Additionally, I followed up PCA by creating a hierarchical clustering heatmap. The heatmap showed very strong correlations (greater than 97%) between most species, expect for humans (Fig. 3). The highest correlation for humans was with chimpanzees. The mouse lemur also demonstrated relatively weaker correlations compared to other non-human primates.

Figure 2: PCA of Expression Patterns Between Non-human Primates and Humans

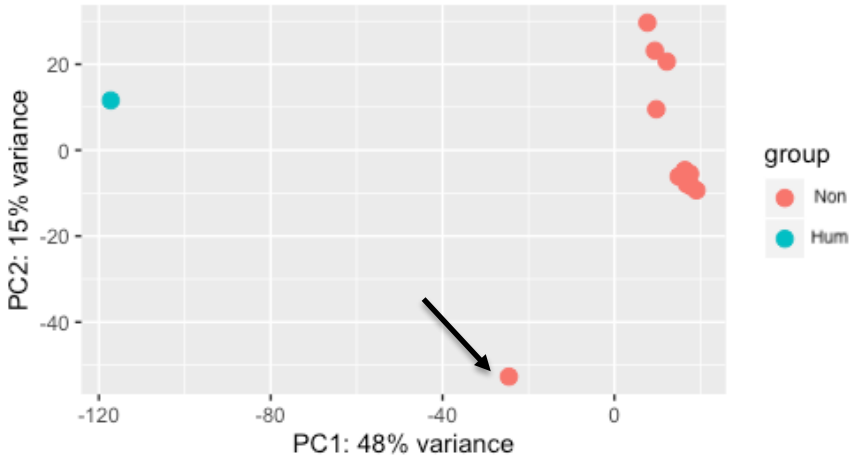


Figure 2: Principal component analysis of expression patterns between non-human primate (Non) and human (Hum) in skeletal muscle. The plot area represents transcriptional space, and each point represents a species in the data set. Humans (blue) and chimpanzees (indicated by arrow) appear to be the two significant outliers.

Figure 3: Heatmap of Gene Expression Correlation Between Primate Species

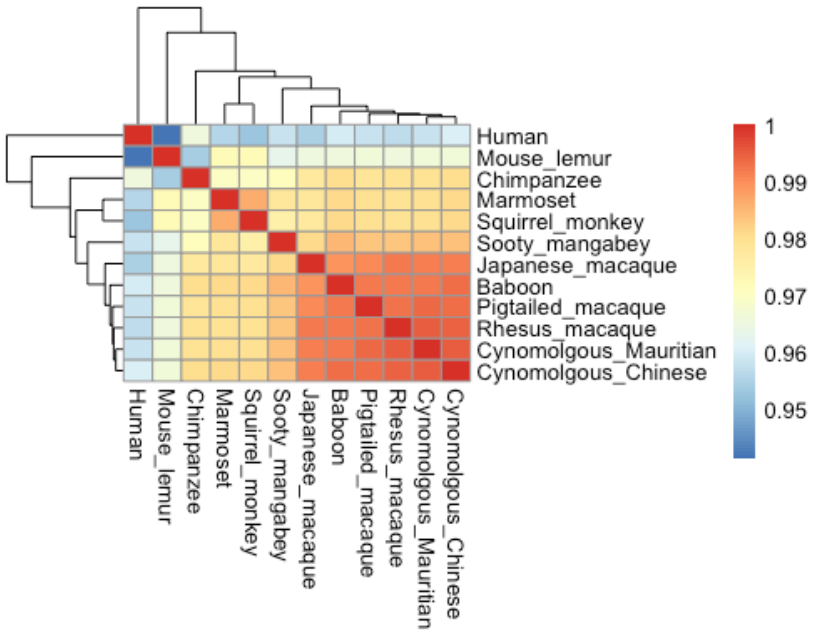


Figure 3: Hierarchical clustering heatmap displaying the correlation in gene expression between each primate species in the data set. The tree demonstrates the projected relatedness based on the expression patterns, and the color boxes indicates strength of correlation between sample.

2,426 genes differentially expressed between humans and non-human primates

Using DESeq2, I conducted a genome-wide survey of all genes differentially expressed in human skeletal muscle tissue relative to other primates. To visualize this list created by this survey, I created a plot of log-fold changes (LFC) for RNA-counts of every gene in the data set (Fig. 4). Out of the 20,420 genes in the data set, 2,426 were differentially expressed. Specifically, 813 genes (4% of the total gene data set) had a significantly positive LFC, meaning those gene are up-regulated in human skeletal muscle fiber relative to non-human primate skeletal muscle fiber (Fig 5 & see Appendix C for the top up-regulated genes). 1,613 genes (7.9% of the total gene list) had a significantly negative LFC, indicating those genes were down-regulated in humans (Fig. 5 & see Appendix C for the top down-regulated genes).

Figure 4: Normalized Gene Expression Counts

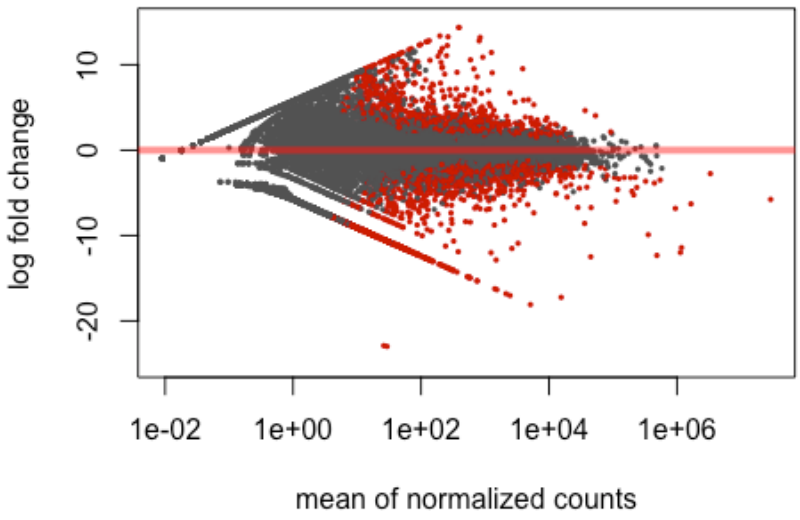


Figure 4: Normalized counts of all expressed genes by their log fold changes. Differentially expressed genes are highlighted red (p-value 0.05).

Figure 5: Differentially Expressed Gene Between Humans and Non-human Primates

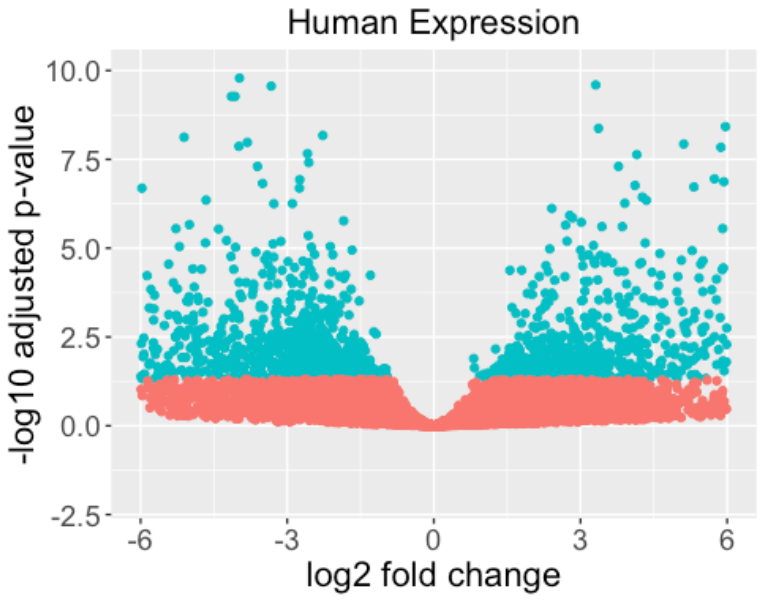


Figure 5: Log fold change values for all differentially expressed genes between humans and other primates plotted against the log 10 adjusted p-value.

Candidate gene expression not significantly different between humans and other primates

My four candidate genes were not among those identified by the survey as differentially expressed (Appendix C and D), so hypothesizes H2A1 and H2A2 were refuted. To visualize these results, I created individual expression plots for each gene. The expression plots for my candidate genes demonstrate that *AOX1*, *FNIP1*, *MYOZ1*, and *PPARG* are not differentially expressed in humans relative to other primates because the human expression point falls within the range of non-human primate expression, supporting hypothesis H2A3 (Fig 6 A-D).

Figure 6: Individual Candidate Gene Expression Plots

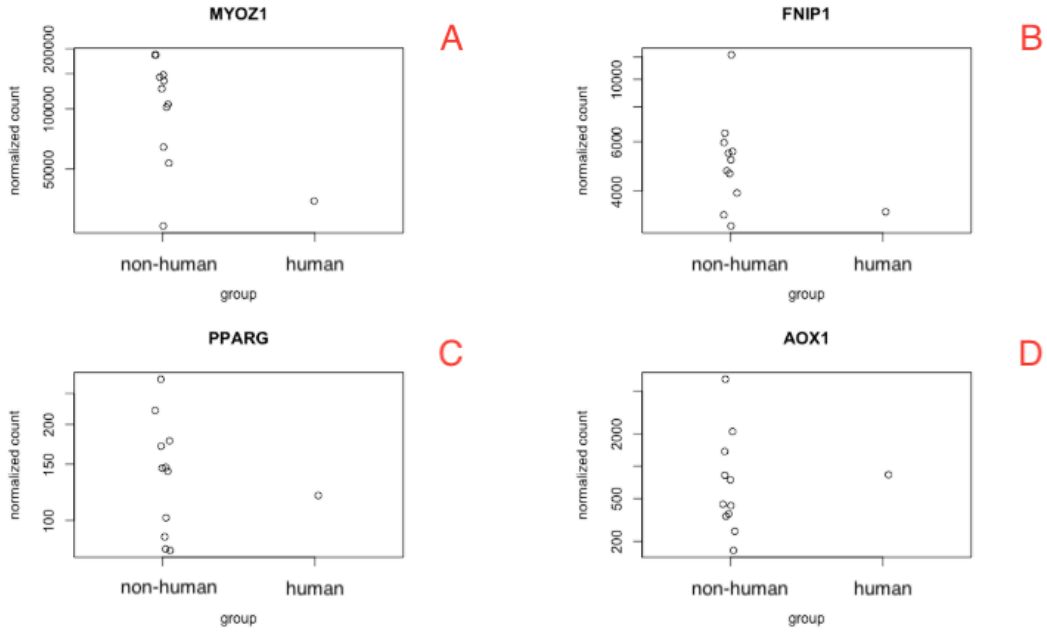


Figure 6: Normalized RNA-counts for MYOZ1 (A), FNIP1 (B), PPARG (C), and AOX1 (D) for non-human primates and humans.

Functional analysis using DAVID

Once I had compiled my lists of up regulated and down regulated genes (cite app table), I used DAVID to analyze which biological processes, cellular components, and molecular functions were over represented in my gene lists. Of the 813 up regulated genes, 669 matched gene IDs in the DAVID database. Many genes that mapped to a biological process involved negative regulation of molecular or larger scale processes (Table 2). The cellular components best represented were related to the cytoplasm, intracellular space, and ribosomes (Table 3). The molecular functions best represented involved binding interactions, including to RNA (Table 4). Of the 1,613 genes down regulated in humans relative to other primates, 1,425 matched DAVID gene IDs. The primary biological processes over represented by this gene list involved cellular or chromosome organization (Table 5). Interestingly, the category of muscle system process appeared in the 10 most significantly represented biological processes, which supports hypothesis H2B1. The main cellular components represented in the down regulated genes included nucleosomes, cytoskeleton, and — like biological processes — chromosomes (Table 6). Another muscle related category also appeared on the top 10 cellular component list: myosin complex. The molecular functions best represented by the down regulated genes included many binding interactions, particularly purine nucleoside/nucleotide binding (Table 7).

Table 2: Overrepresented Biological Processes in Up-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_BP_3	negative regulation of programmed cell death	RT		29	4.3	1.1E-4	7.1E-2
GOTERM_BP_3	negative regulation of cell death	RT		29	4.3	1.2E-4	3.8E-2
GOTERM_BP_3	negative regulation of nitrogen compound metabolic process	RT		36	5.4	2.9E-4	6.1E-2
GOTERM_BP_3	regulation of binding	RT		16	2.4	4.3E-4	6.8E-2
GOTERM_BP_3	intracellular transport	RT		42	6.3	4.3E-4	5.5E-2
GOTERM_BP_3	negative regulation of biosynthetic process	RT		38	5.7	4.4E-4	4.8E-2
GOTERM_BP_3	peptide metabolic process	RT		9	1.3	5.8E-4	5.3E-2
GOTERM_BP_3	negative regulation of metabolic process	RT		47	7.0	7.0E-4	5.6E-2
GOTERM_BP_3	negative regulation of cellular metabolic process	RT		44	6.6	8.0E-4	5.8E-2
GOTERM_BP_3	negative regulation of biological process	RT		91	13.6	8.6E-4	5.6E-2

Table 3: Overrepresented Cellular Components in Up-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_CC_3	cytoplasm	RT		346	51.7	6.2E-15	9.9E-13
GOTERM_CC_3	intracellular part	RT		439	65.6	7.6E-11	6.0E-9
GOTERM_CC_3	intracellular	RT		449	67.1	1.7E-10	8.8E-9
GOTERM_CC_3	cytoplasmic part	RT		240	35.9	5.1E-10	2.0E-8
GOTERM_CC_3	ribosomal subunit	RT		20	3.0	1.1E-7	3.4E-6
GOTERM_CC_3	intracellular organelle	RT		372	55.6	3.0E-7	8.0E-6
GOTERM_CC_3	ribosome	RT		24	3.6	2.1E-6	4.7E-5
GOTERM_CC_3	intracellular organelle part	RT		195	29.1	7.9E-6	1.6E-4
GOTERM_CC_3	intracellular membrane-bounded organelle	RT		330	49.3	1.1E-5	1.9E-4
GOTERM_CC_3	ribonucleoprotein complex	RT		39	5.8	1.3E-5	2.1E-4

Table 4: Overrepresented Molecular Functions in Up-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_MF_3	RNA binding	RT		44	6.6	1.4E-4	1.9E-2
GOTERM_MF_3	transcription factor binding	RT		34	5.1	2.3E-4	1.6E-2
GOTERM_MF_3	peptide transporter activity	RT		4	0.6	3.7E-3	1.6E-1
GOTERM_MF_3	cytoskeletal protein binding	RT		29	4.3	5.7E-3	1.8E-1
GOTERM_MF_3	TAP binding	RT		3	0.4	6.4E-3	1.6E-1
GOTERM_MF_3	protein domain specific binding	RT		21	3.1	7.6E-3	1.6E-1
GOTERM_MF_3	protein dimerization activity	RT		30	4.5	8.3E-3	1.5E-1
GOTERM_MF_3	peptide antigen binding	RT		4	0.6	1.2E-2	1.9E-1
GOTERM_MF_3	translation initiation factor activity	RT		7	1.0	1.6E-2	2.2E-1
GOTERM_MF_3	transcription corepressor activity	RT		11	1.6	2.3E-2	2.8E-1

Table 7-9: DAVID functional analysis results for up regulated genes in human skeletal muscle tissue. Categories are grouped by biological process (7), cellular component (8) and molecular function (9).

Table 5: Overrepresented Biological Processes in Down-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_BP_3	chromatin assembly	RT		24	1.7	3.4E-8	2.5E-5
GOTERM_BP_3	nucleosome organization	RT		22	1.5	2.3E-6	8.4E-4
GOTERM_BP_3	microtubule-based movement	RT		23	1.6	1.7E-5	4.2E-3
GOTERM_BP_3	chromosome organization	RT		58	4.1	1.9E-4	3.4E-2
GOTERM_BP_3	muscle system process	RT		25	1.8	1.0E-3	1.4E-1
GOTERM_BP_3	regulation of organelle organization	RT		29	2.0	2.0E-3	2.2E-1
GOTERM_BP_3	ion transport	RT		77	5.4	3.1E-3	2.8E-1
GOTERM_BP_3	cell cycle phase	RT		46	3.2	4.1E-3	3.2E-1
GOTERM_BP_3	mitotic cell cycle	RT		41	2.9	7.0E-3	4.4E-1
GOTERM_BP_3	organelle fission	RT		28	2.0	8.3E-3	4.6E-1

Table 6: Overrepresented Cellular Components in Down-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_CC_3	nucleosome	RT		19	1.3	1.7E-7	3.3E-5
GOTERM_CC_3	cytoskeletal part	RT		107	7.5	7.8E-7	7.7E-5
GOTERM_CC_3	chromosomal part	RT		54	3.8	1.6E-6	1.0E-4
GOTERM_CC_3	intracellular non-membrane-bounded organelle	RT		234	16.4	1.2E-5	5.8E-4
GOTERM_CC_3	myosin complex	RT		16	1.1	3.1E-5	1.2E-3
GOTERM_CC_3	cell projection	RT		76	5.3	1.1E-4	3.5E-3
GOTERM_CC_3	extracellular matrix part	RT		19	1.3	1.2E-3	3.3E-2
GOTERM_CC_3	basal plasma membrane	RT		8	0.6	1.6E-3	3.8E-2
GOTERM_CC_3	extracellular matrix	RT		40	2.8	1.9E-3	4.1E-2
GOTERM_CC_3	kinetochore	RT		14	1.0	2.4E-3	4.6E-2

Table 7: Overrepresented Molecular Functions in Down-regulated Gene List

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_MF_3	purine nucleoside binding	RT		171	12.0	9.8E-8	1.7E-5
GOTERM_MF_3	cation binding	RT		378	26.5	1.1E-7	9.6E-6
GOTERM_MF_3	ribonucleotide binding	RT		177	12.4	3.8E-5	2.3E-3
GOTERM_MF_3	purine nucleotide binding	RT		183	12.8	4.8E-5	2.1E-3
GOTERM_MF_3	hydrolase activity, acting on acid anhydrides	RT		83	5.8	2.0E-4	7.2E-3
GOTERM_MF_3	transferase activity, transferring phosphorus-containing groups	RT		100	7.0	2.3E-4	6.8E-3
GOTERM_MF_3	DNA binding	RT		207	14.5	9.3E-4	2.3E-2
GOTERM_MF_3	cytoskeletal protein binding	RT		54	3.8	4.1E-3	8.7E-2
GOTERM_MF_3	passive transmembrane transporter activity	RT		46	3.2	4.1E-3	7.8E-2
GOTERM_MF_3	calmodulin binding	RT		19	1.3	1.3E-2	2.0E-1

Tables 5-7: DAVID functional analysis results for down regulated genes in human skeletal muscle tissue. Categories are grouped by biological process (5), cellular component (6) and molecular function (7).

Discussion

Lack of evidence for genetic variation in genes related to skeletal muscle fiber type.

Through my first objective, I wanted to explore if the genetic sequences of genes related to skeletal muscle fiber type were conserved across primate species. I also investigated if such genes were under positive selection. To narrow my focus, I selected *AOX1*, *FNIP1*, *MYOZ1* and *PPARG* as candidate genes for my analysis. Since fiber construction is generally conserved across primates and mammals in general, I hypothesized that the candidate sequences would be conserved across species and that purifying selection likely selects against new variants in these genes. I found little variation in structure of the four candidate gene sequences, which suggests genetic variation is not influencing skeletal muscle fiber type. This sequence conservation is likely because changing the structure of DNA sequences can change — sometimes drastically — the protein produced by the sequence. Besides the length of fibers being shorter on average in humans (O'Neill 2017), the structure of skeletal muscle fibers themselves do not vary between humans and other primates — only the relative abundance of slow to fast twitch fibers in each species varies. I found evidence of positive selection on *AOX1* and *FNIP1*, though, no individual sites in *AOX1* and just one site in *FNIP1* were specifically identified. Further research is needed to assess the significance of these findings. Despite this result, purifying selection, or the selection against deleterious mutations is probably a greater influence on the candidate genes given the sequence conservation and lack of significant positive selection sites.

My main limitation for objective 1 was identifying appropriate candidate loci when so little is known about the molecular mechanisms of fiber type. Four genes are not enough to make inferences about all gene sequences relevant to skeletal muscle fiber. As is the case with many human and other primate functions, skeletal muscle fiber composition might be influenced by several genes, all of which might vary between species. A much larger sample size of genes would be needed to detect a wide array of influences. Additionally, the fact that *AOX1* appears to be under positive selection yet no regions of the sequence were deemed significantly under positive selection by BEB analysis means the detection of *AOX1* could be a false positive. This may suggest this finding was a false positive or the signal was too low to detect specific sites under positive selection.

Differentially expressed genes exist in skeletal muscle, but further research required

I also wanted to test if my four candidate genes are differentially expressed in primate skeletal muscle tissues. I hypothesized that my four candidate genes — *AOX1*, *FNIP1*, *MYOZ1* and *PPARG* — were differentially expressed in humans compared to non-human primates given the genes' abilities to control skeletal muscle fiber ratio in mice (Chemello 2011, Wang 2004, Reyes 2015). The data I collected refuted this hypothesis: the candidate genes were not identified as differentially expressed in human skeletal muscle tissue. However, I was able to identify many other up-regulated and down-regulated genes in humans relative to other primates. This still suggests that regulatory differences in skeletal muscle related genes may have the largest effect on the ratio of slow-twitch to fast-twitch fibers. I was also able to determine that several genes in the down-regulated list had been linked to some aspect of skeletal muscle fiber

before. Further refinement of this list is necessary to parse out which genes are being down-regulated in human skeletal muscle fiber types specifically, as those genes might be responsible for determining whether a skeletal muscle fiber becomes slow or fast-twitch.

My DESeq2 results should be interpreted with some caution, as there were several limitations I encountered. First, the Non-Human Primate Reference Transcriptome Resource NHPRTR does not list from which specific muscle their samples were acquired from in each species. While I have referred to skeletal muscle fiber ratio as an average across species, the ratio varies greatly between muscles in the same species, particularly from the lower limbs to the upper limbs. A sample of muscle from a human bicep would have a much greater percentage of fast-twitch fibers relative to a sample from a human quadricep. The RNA-seq data produced from sampling a bicep might not be indicative of the genetic expression as it pertains to the average fiber ratio. Second, DESeq2 works best when comparing two groups of species' genetic expression. DESeq2 is trying to determine for each gene whether the differences in expression (counts) between groups is significant given the amount of variation observed within groups (replicates).When one species is isolated — as humans are in my comparison — the expression of that species may become artificially inflated relative to the comparison group (non-human primates), as there are no replicates to compare to.

However, none of the other primates in the database had a greater percentage of slow-twitch fiber, so it would have been inappropriate to group humans with any of the other primates given the question I was interested in. If slow loris transcriptome data

was available to me, that would have greatly strengthened my comparison, as they are one of the few non-human primates with more slow-twitch fibers. Third, chimpanzees are the only other ape present in the sample group, so the data disproportionally represents non-ape primates. Had I an opportunity to conduct a follow-up study, I would pursue two objectives. First, I would sample the gluteus maximus of slow lorises, humans, multiple ape species, and multiple Old World and New World monkey species. The gluteus maximus is one of the primary muscles involved in bipedal locomotion, and all primates have a gluteus maximus muscle or the close equivalent (Janković 2015). This study design would enable me to group together humans and slow lorises to then compare to the rest of the non-human primates. Genes identified as differentially expressed as a result of this hypothetical study would be more likely to be directly linked to fiber ratio than the genes in this study — which may be specific to humans for reasons other than fiber ratio. As a second objective of this study, I would sample transcriptome data from individual slow- and fast-twitch fibers in humans and another apes. This would allow me to distinguish which genes are up- or down-regulated in each type of fiber. Ideally, I could cross-reference the genes differentially expressed in the first objective with genes identified in the second objective: genes identified in both objectives would be compelling candidates for determinants of skeletal muscle fiber type.

Broader Impacts and Beyond

The results from this project will contribute to an understudied area of human evolution. Skeletal muscle does not receive the same attention other aspects of our evolutionary history do — chiefly because muscles cannot fossilize — but muscles are

critical to consider in the overall story of how modern humans evolved. Understanding the *origins* of human bipedalism is only half the story: how humans maximized the efficiency of bipedal locomotion completes the narrative. Through this project, I have compiled a list of differentially expressed genes in human skeletal muscle tissue relative to other primates. These genes could make for suitable candidate genes for future studies on skeletal muscle.

In working through the second objective of this project, I found publicly available transcriptome data to be quite limited. The National Human Primate Reference Transcriptome Resource contains just 13 species including humans with data specific to skeletal muscle fiber, compared to the 24 species available in genome browsers Furthermore, the muscle of origin from these samples is unknown. As skeletal muscle fiber ratio varies between muscles in an individual, identifying which muscle a sample is taken from is critical information to consider when making conclusions about differential gene expression. Fiber type-specific samples do not exist either. Had I access to transcriptome data from individual slow-twitch fibers and fast twitch fibers, I would have been able to make more robust conclusions about genes differentially expressed between fiber types. I hope that my critiques will motivate action to add samples from additional species and tissue types to this database.

This project has medical applications as well. A better understanding of genes at play in muscle fiber regulation might inform those working on potential gene therapy treatments for patients dealing with muscle related diseases. Wasting or atrophy of muscle might be mediated or reversed by activation or repression of a particular gene, so figuring out which genes are active in which muscle fiber type would help narrow

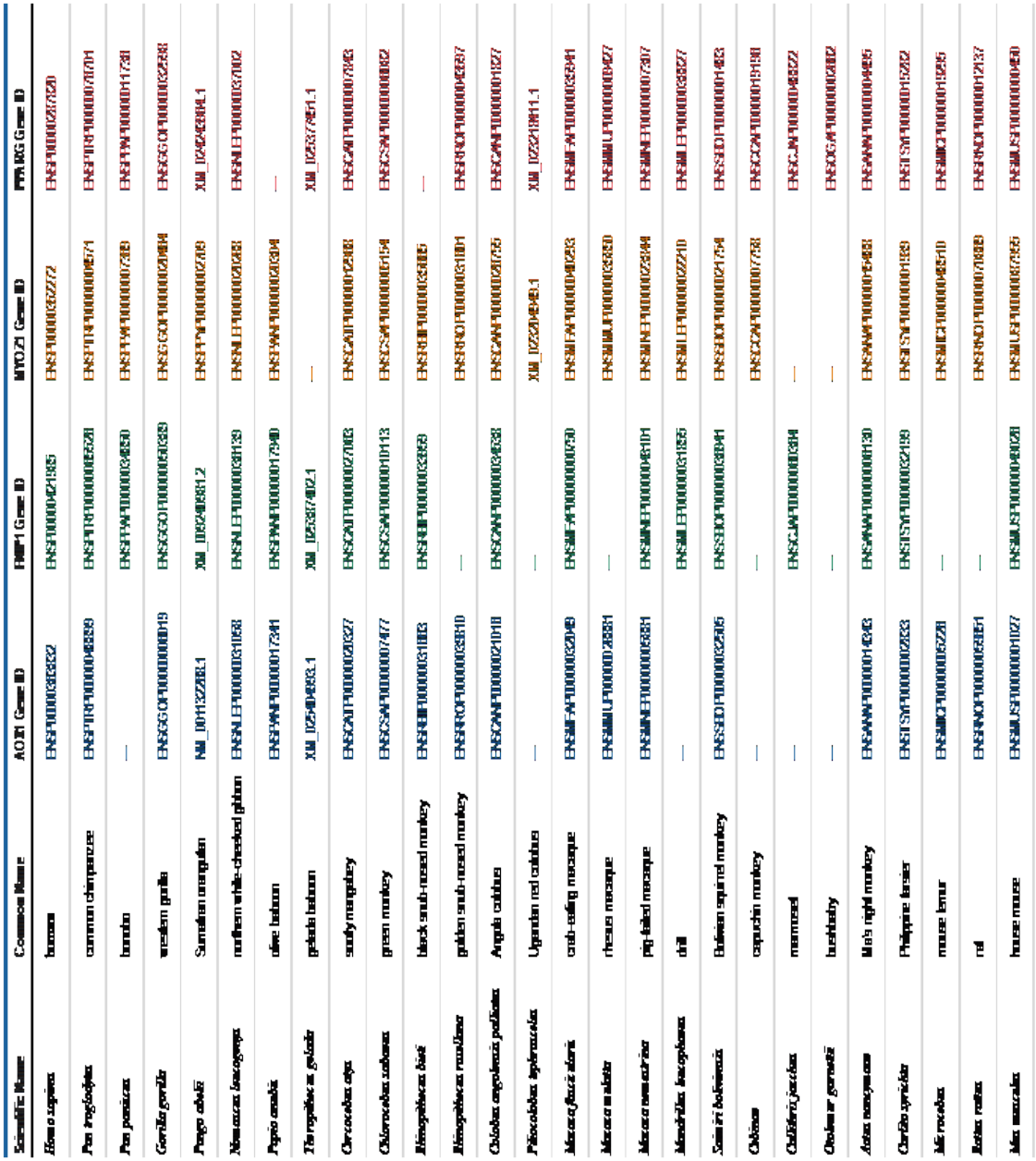
the search for new therapies. Some research has also found an association between loss of muscle function and increased risk of developing type II diabetes (Kelley 2002), so regulating gene expression of muscle fibers would be useful in that case as well.

Conclusions

Bipedalism is a defining human characteristic. Over the course of human evolution, human skeletal structure has changed in a variety of ways to accommodate to this locomotion strategy (Ko 2015). But while extensive research has addressed the origins of bipedalism using fossil evidence, less research been conducted on how humans maximized bipedal efficiency. One way locomotive efficiency has been increased in humans is through increasing the ratio of slow-twitch to fast-twitch fibers in skeletal muscle. Human skeletal muscle fiber contains more slow-twitch fibers than most all other primates (O'Neil 2017). Since this trait seems so particular to humans, I wanted to investigate the source of this ratio difference. Specifically, I wanted to know if genetic variation or differential gene expression in humans was responsible for the fiber type ratio difference. Ultimately, I found little evidence genetic variation within protein-coding genes was influencing fiber type ratio. I did identify many genes differentially expressed between human skeletal muscle and other primate skeletal muscle tissue, but further research is required before any of those genes are confirmed to influence skeletal fiber type. The goal of the project was to establish initial inquiry into the subject using the public resources available to me, so while I was not able to comprehensively answer my research question, I hope my project will inspire further research on this matter.

Appendix

Appendix A: Species List and Gene ID Codes



Scientific Name	Common Name	AOX1 Gene ID	FMO1 Gene ID	MYO7A Gene ID	PPARGC1A Gene ID
Homo sapiens	human	ENSP00000352552	ENSP00000421935	ENSP00000352772	ENSP00000267020
Pan troglodytes	common chimpanzee	ENSPTRP00000045559	ENSPTRP00000055281	ENSPTRP00000055474	ENSPTRP00000070701
Pan paniscus	bonobo	—	ENSPPAP00000034850	ENSPPAP00000007380	ENSPPAP00000041731
Gorilla gorilla	western gorilla	ENSGGOP00000000019	ENSGGOP00000030389	ENSGGOP00000029814	ENSGGOP00000032288
Pongo abelii	Sumatran orangutan	NM_001132728.1	XM_002828812.2	ENSPPYP00000007389	XM_024642894.1
Nomascus leucogenys	northern white-cheeked gibbon	ENSNLEP00000030938	ENSNLEP00000030938	ENSNLEP00000026581	ENSNLEP00000037002
Papio anubis	olive baboon	ENSPANP00000017341	ENSPANP00000017940	ENSPANP00000026304	—
Theropithecus gelada	gelada baboon	XM_025049833.1	XM_025057402.1	—	XM_025377851.1
Cercocebus atys	sooty mangabey	ENSCATP00000028227	ENSCATP00000027063	ENSCATP00000025881	ENSCATP00000037043
Chlorocebus sabaeus	green monkey	ENSCSAP00000007477	ENSCSAP00000010113	ENSCSAP00000006154	ENSCSAP00000000852
Rhinopithecus bieti	black snub-nosed monkey	ENSRBIP00000031883	ENSRBIP00000032858	ENSRBIP00000035805	—
Rhinopithecus roxellana	golden snub-nosed monkey	ENSRROP00000038810	—	ENSRROP00000031001	ENSRROP00000042837
Colobus angolensis palliatus	Angola colobus	ENSCANP00000021018	ENSCANP00000034838	ENSCANP00000021655	ENSCANP00000031827
Piliocolobus tephrosceles	Ugandan red colobus	—	—	XM_023204045.1	XM_023210811.1
Macaca fascicularis	crab-eating macaque	ENSMFAP00000032943	ENSMFAP00000030750	ENSMFAP00000048283	ENSMFAP00000032941
Macaca mulatta	rhesus macaque	ENSMMUP00000028531	—	ENSMMUP00000035350	ENSMMUP00000004827
Macaca nemestrina	pig-tailed macaque	ENSMNEP00000005831	ENSMNEP00000046101	ENSMNEP00000025044	ENSMNEP00000007307
Mandrillus leucophaeus	drill	—	ENSMLEP00000031855	ENSMLEP00000022210	ENSMLEP00000038527
Saimiri boliviensis	Bolivian squirrel monkey	ENSSBOP00000032505	ENSSBOP00000036941	ENSSBOP00000021754	ENSSBOP00000001493
Cebus	capuchin monkey	—	—	ENSCCAP00000007758	ENSCCAP00000019181
Callithrix jacchus	marmoset	—	ENSCJAP00000030814	—	ENSCJAP00000048522
Cebus capucinus imitator	white-faced capuchin	—	—	—	ENSCCAP00000028182
Aotus nancymaae	black-headed night monkey	ENSANAP00000014343	ENSANAP00000000130	ENSANAP00000015453	ENSANAP00000004825
Carlito syrichta	Philippine tarsier	ENSTSYP00000002553	ENSTSYP00000032189	ENSTSYP00000001189	ENSTSYP00000015282
Microcebus	mouse lemur	ENSMICP00000035228	—	ENSMICP00000035510	ENSMICP00000019255
Rattus rattus	rat	ENSRNOP00000058861	—	ENSRNOP00000070888	ENSRNOP00000012137
Mus musculus	house mouse	ENSMUSP00000011027	ENSMUSP00000048281	ENSMUSP00000037855	ENSMUSP00000000450

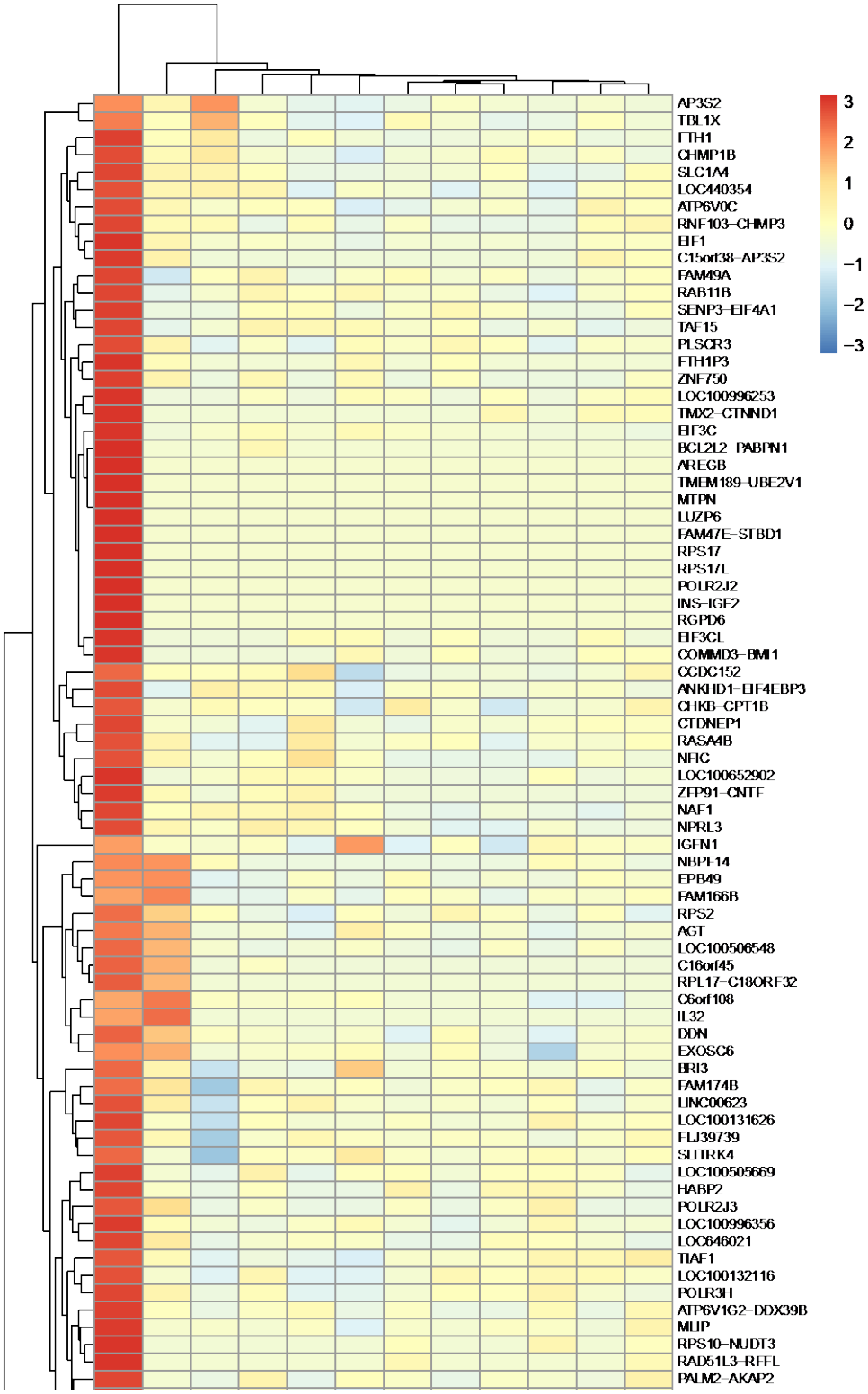
Appendix A: Scientific name and common name of all species included in genetic variation analysis. Gene IDs refer to the specific sequence acquired from Ensembl or NCBI Blast. Dashes denote where a species was not included in the analysis as the sequence was incomplete or corrupted.

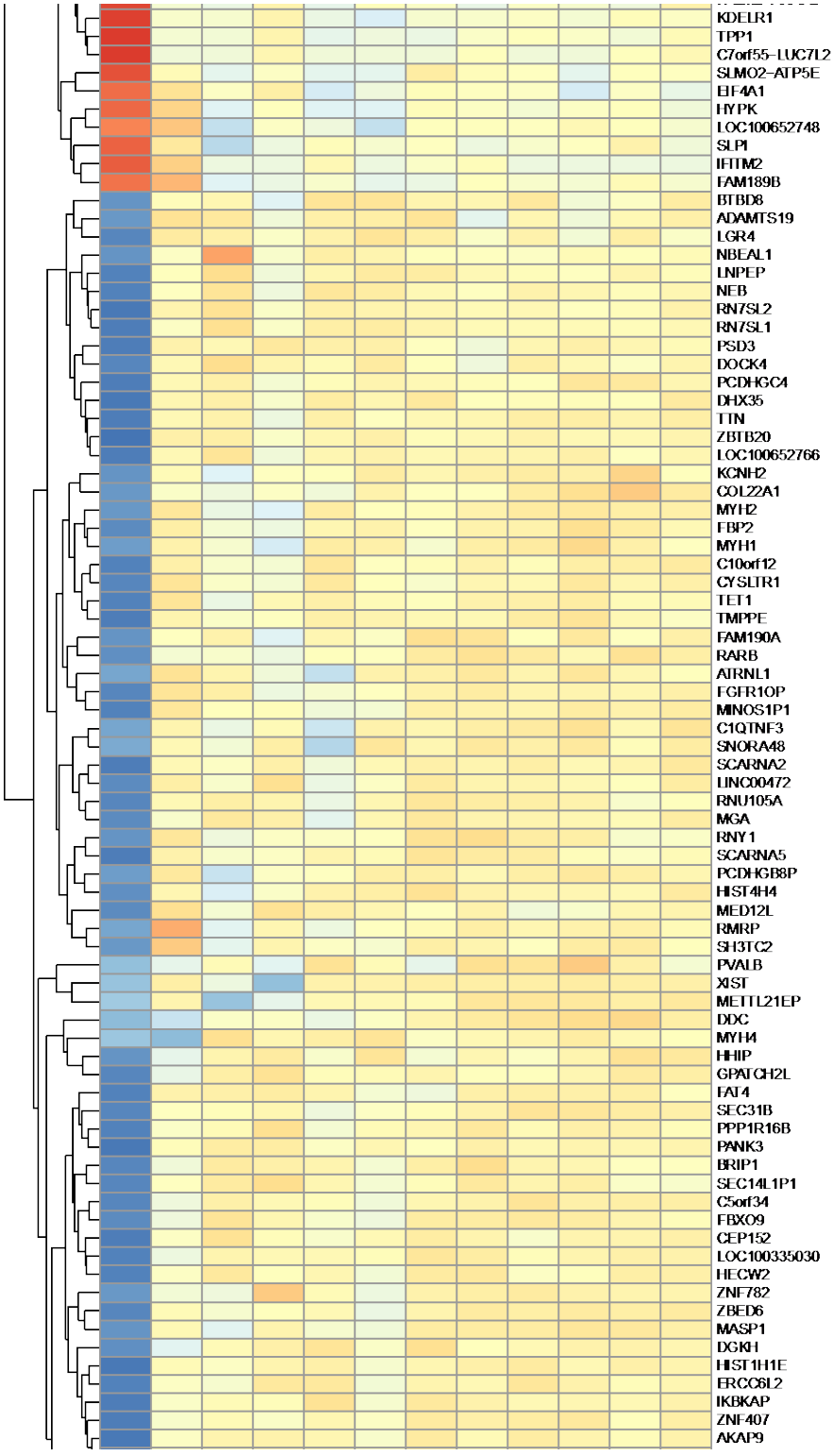
Appendix B: Metadata Table Used in DESeq2

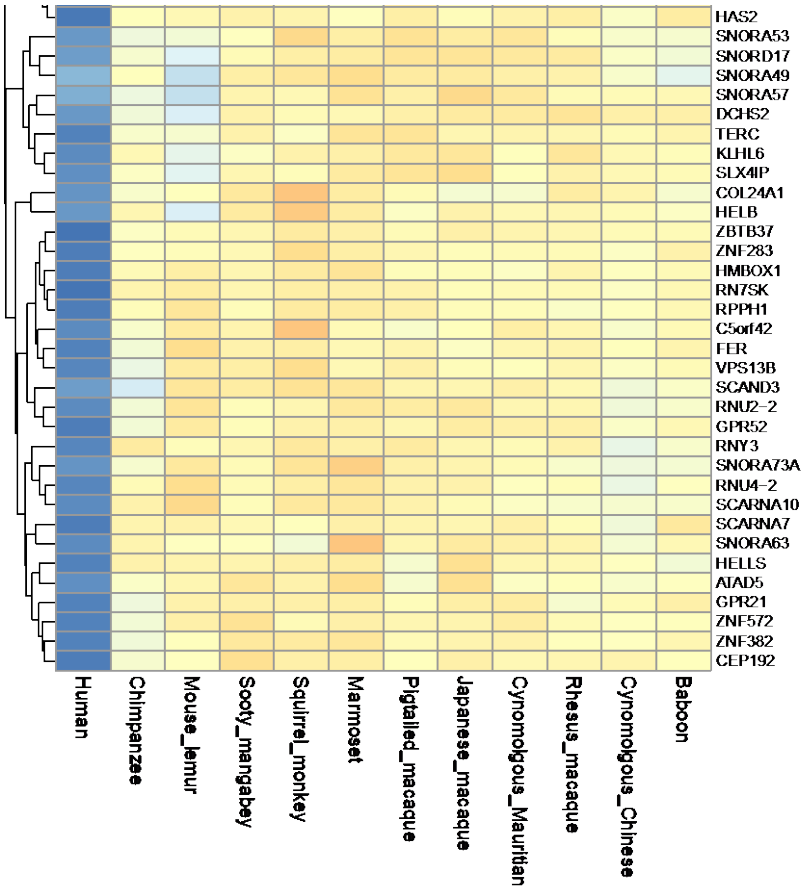
Species Name	Human (Y or N)	Non-human (Y or N)
human	Y	N
common chimpanzee	N	Y
pig-tailed macaque	N	Y
Japanese macaque	N	Y
rhesus macaque	N	Y
crab-eating macaque (Mauritian)	N	Y
crab-eating macaque (Chinese)	N	Y
olive baboon	N	Y
sooty mangabey	N	Y
common marmoset	N	Y
squirrel monkey	N	Y
mouse emur	N	Y

Appendix B: Metadata read by DESeq2 that specifies the groupings during the analysis. A "Y" means that species was included in the given group (Human or Non-human), a "N" indicates the species was not included in the group.

Appendix C: Differentially Expressed Gene List (partial)




KDELR1
TPP1
C7orf55-LUC7L2
SUMO2-ATP5E
EIF4A1
HYPK
LOC100652748
SLPI
IFITM2
FAM189B
BTBD8
ADAMTS19
LGR4
NBEAL1
LNPEP
NEB
RN7SL2
RN7SL1
PSD3
DOCK4
PCDHGC4
DHX35
TTN
ZBTB20
LOC100652766
KCNH2
COL22A1
MYH2
FBP2
MYH1
C10orf12
CYSLTR1
TET1
TMPPE
FAM190A
RARB
ATRNL1
FGFR1OP
MINOS1P1
C1QTNF3
SNORA48
SCARNA2
LINC00472
RNU105A
MGA
RNY1
SCARNA5
PCDHGB8P
HIST4H4
MED12L
RMRP
SH3TC2
PVALB
XIST
METTL21EP
DDC
MYH4
HHIP
GPATCH2L
FAT4
SEC31B
PPP1R16B
PANK3
BRIP1
SEC14L1P1
C5orf34
FBXO9
CEP152
LOC100335030
HECW2
ZNF782
ZBED6
MASP1
DGKH
HIST1H1E
ERCC6L2
IKBKAP
ZNF407
AKAP9



Appendix C: Heatmap showing the list of the top 188 differentially expressed genes in human skeletal muscle fiber identified by DESeq2 (p-value < 0.0001). Z-score scale indicates the degree of differential expression (red = up-regulated, blue = down-regulated).

Bibliography

Almécija, S., Tallman, M., Alba, D. M., Pina, M., Moyà-Solà, S., & Jungers, W. L. (2013, 12). The femur of Orrorin tugenensis exhibits morphometric affinities with both Miocene apes and later hominins. *Nature Communications, 4*(1). doi:10.1038/ncomms3888

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990, 10). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Baba, M., Hong, S., Sharma, N., Warren, M. B., Nickerson, M. L., Iwamatsu, A., . . . Zbar, B. (2006, 10). Folliculin encoded by the BHD gene interacts with a binding protein, FNIP1, and AMPK, and is involved in AMPK and mTOR signaling. *Proceedings of the National Academy of Sciences, 103*(42), 15552-15557. doi:10.1073/pnas.0603781103

Barber, R. D., Harmer, D. W., Coleman, R. A., & Clark, B. J. (2005, 05). GAPDH as a housekeeping gene: Analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiological Genomics, 21*(3), 389-395. doi:10.1152/physiolgenomics.00025.2005

Bramble, D. M., & Lieberman, D. E. (2004, 11). Endurance running and the evolution of Homo. *Nature, 432*(7015), 345-352. doi:10.1038/nature03052

Carbone, C., Cowlishaw, G., Isaac, N., & Rowcliffe, J. (2005, 02). How Far Do Animals Go? Determinants of Day Range in Mammals. *The American Naturalist, 165*(2), 290-297. doi:10.1086/426790

Carrier, D. R., Kapoor, A. K., Kimura, T., Nickels, M. K., Scott, E. C., So, J. K., & Trinkaus, E. (1984, 08). The Energetic Paradox of Human Running and Hominid Evolution [and Comments and Reply]. *Current Anthropology, 25*(4), 483-495. doi:10.1086/203165

Chemello, F., Bean, C., Cancellara, P., Laveder, P., Reggiani, C., & Lanfranchi, G. (2011, 02). Microgenomic Analysis in Skeletal Muscle: Expression Signatures of Individual Fast and Slow Myofibers. *PLoS ONE, 6*(2). doi:10.1371/journal.pone.0016807

Chojnacki, S., Cowley, A., Lee, J., Foix, A., & Lopez, R. (2017, 04). Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research, 45*(W1). doi:10.1093/nar/gkx273

Frontera, W. R., & Ochala, J. (2014, 10). Skeletal Muscle: A Brief Review of Structure and Function. *Calcified Tissue International, 96*(3), 183-195. doi:10.1007/s00223-014-9915-y

Garattini, E., Fratelli, M., and Terao, M. (2009). The mammalian aldehyde oxidase gene family. *Human genomics* 4(2): 119-30.

Green, D. J., Gordon, A. D., & Richmond, B. G. (2007, 02). Limb-size proportions in Australopithecus afarensis and Australopithecus africanus. *Journal of Human Evolution, 52*(2), 187-200. doi:10.1016/j.jhevol.2006.09.001

Haile-Selassie, Y. (2010, 10). Phylogeny of early Australopithecus: New fossil evidence from the Woranso-Mille (central Afar, Ethiopia). *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1556), 3323-3331. doi:10.1098/rstb.2010.0064

Hatala, K. G., Wunderlich, R. E., Dingwall, H. L., & Richmond, B. G. (2016, 01). Interpreting locomotor biomechanics from the morphology of human footprints. *Journal of Human Evolution, 90*, 38-48. doi:10.1016/j.jhevol.2015.08.009

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008, 12). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57. doi:10.1038/nprot.2008.211

Hunt, K. D. (1994, 03). The evolution of human bipedality: Ecology and functional morphology. *Journal of Human Evolution, 26*(3), 183-202. doi:10.1006/jhev.1994.1011

Janković, Ivor. (2015). Certain medical problems resulting from evolutionary processes: Bipedalism as an example. *Periodicum Biologorum*. 117.

Kelley, D. E., He, J., Menshikova, E. V., & Ritov, V. B. (2002, 10). Dysfunction of Mitochondria in Human Skeletal Muscle in Type 2 Diabetes. *Diabetes, 51*(10), 2944-2950. doi:10.2337/diabetes.51.10.2944

Kimbel, W. H., Suwa, G., Asfaw, B., Rak, Y., & White, T. D. (2014, 01). Ardipithecus ramidus and the evolution of the human cranial base. *Proceedings of the National Academy of Sciences, 111*(3), 948-953. doi:10.1073/pnas.1322639111

Klitgaard, H., Mantoni, M., Schiaffino, S., Ausoni, S., Gorza, L., Laurent-Winter, C., . . . Saltin, B. (1990, 09). Function, morphology and protein expression of ageing skeletal muscle: A cross-sectional study of elderly men with different training backgrounds. *Acta Physiologica Scandinavica, 140*(1), 41-54. doi:10.1111/j.1748-1716.1990.tb08974.x

Ko, K. H. (2015, 12). Origins of Bipedalism. *Brazilian Archives of Biology and Technology, 58*(6), 929-934. doi:10.1590/s1516-89132015060399

Kuliukas, A. (2002, 10). Wading for Food the Driving Force of the Evolution of Bipedalism? *Nutrition and Health, 16*(4), 267-289. doi:10.1177/026010600201600402

Kumar, S., Stecher, G., & Tamura, K. (2016, 03). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution, 33*(7), 1870-1874. doi:10.1093/molbev/msw054

Li, Y., Sun, L., Shi, Y., Wang, G., Wang, X., Dunn, S. E., . . . Spaner, D. E. (2017, 01). PPAR-delta promotes survival of chronic lymphocytic leukemia cells in energetically unfavorable conditions. *Leukemia, 31*(9), 1905-1914. doi:10.1038/leu.2016.395

Lieberman, D. E. (2014, 12). Human Locomotion and Heat Loss: An Evolutionary Perspective. *Comprehensive Physiology,* 99-117. doi:10.1002/cphy.c140011

Lin, H., Dolmatova, E. V., Morley, M. P., Lunetta, K. L., Mcmanus, D. D., Magnani, J. W., . . . Ellinor, P. T. (2014, 02). Gene expression and genetic variation in human atria. *Heart Rhythm, 11*(2), 266-271. doi:10.1016/j.hrthm.2013.10.051

Lovejoy, C. O. (1988, 11). Evolution of Human Walking. *Scientific American, 259*(5), 118-125. doi:10.1038/scientificamerican1188-118

Lovejoy, C. O. (2007, 03). The natural history of human gait and posture. *Gait & Posture, 25*(3), 325-341. doi:10.1016/j.gaitpost.2006.05.001

Maddison, W. P. and D.R. Maddison. 2018. Mesquite: a modular system for evolutionary analysis. Version 3.51 http://www.mesquiteproject.org

Naya, F. J., Mercer, B., Shelton, J., Richardson, J. A., Williams, R. S., & Olson, E. N. (2000, 02). Stimulation of Slow Skeletal Muscle Fiber Gene Expression by Calcineurinin Vivo. *Journal of Biological Chemistry, 275*(7), 4545-4548. doi:10.1074/jbc.275.7.4545

Neaux, D., Bienvenu, T., Guy, F., Daver, G., Sansalone, G., Ledogar, J. A., . . . Brunet, M. (2017, 12). Relationship between foramen magnum position and locomotion in extant and extinct hominoids. Journal of Human Evolution, 113, 1-9. doi:10.1016/j.jhevol.2017.07.009

Niemitz, C. (2002). A Theory on the Evolution of the Habitual Orthograde Human Bipedalism — The "Amphibische Generalistentheorie". *Anthropologischer Anzeiger*, 60(1), 3-66.

Nekaris, K. (2001). Activity Budget and Positional Behavior of the Mysore Slender Loris (Loris tardigradus lydekkerianus*)*: Implications for Slow Climbing Locomotion. *Folia Primatologica, 72*(4), 228-241. doi:10.1159/000049942

O'Neill, M. C., Umberger, B. R., Holowka, N. B., Larson, S. G., & Reiser, P. J. (2017, 06). Chimpanzee super strength and human skeletal muscle evolution. *Proceedings of the National Academy of Sciences, 114*(28), 7343-7348. doi:10.1073/pnas.1619071114

Pablos, A., Lorenzo, C., Martínez, I., Castro, J. M., Martinón-Torres, M., Carbonell, E., & Arsuaga, J. L. (2012, 10). New foot remains from the Gran Dolina-TD6 Early Pleistocene site (Sierra de Atapuerca, Burgos, Spain). *Journal of Human Evolution, 63*(4), 610-623. doi:10.1016/j.jhevol.2012.06.008

Peng, X., Thierry-Mieg, J., Thierry-Mieg, D., Nishida, A., Pipes, L., Bozinoski, M., . . . Mason, C. E. (2014, 11). Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic Acids Research, 43*(D1). doi:10.1093/nar/gku1110

Peter, J. B., Barnard, R. J., Edgerton, V. R., Gillespie, C. A., & Stempel, K. E. (1972, 07). Metabolic profiles of three fiber types of skeletal muscle in guinea pigs and rabbits. *Biochemistry, 11*(14), 2627-2633. doi:10.1021/bi00764a013

Pipes, L., Li, S., Bozinoski, M., Palermo, R., Peng, X., Blood, P., . . . Katze, M. G. (2012, 11). The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Research, 41*(D1). doi:10.1093/nar/gks1268

Pontzer, H. (2012, 12). Ecological Energetics in Early Homo. *Current Anthropology, 53*(S6). doi:10.1086/667402

Pontzer, H. (2017, 06). Economy and Endurance in Human Evolution. *Current Biology, 27*(12). doi:10.1016/j.cub.2017.05.031

Reyes, N. L., Banks, G. B., Tsang, M., Margineantu, D., Gu, H., Djukovic, D., . . . Iritani, B. M. (2014, 12). Fnip1 regulates skeletal muscle fiber type specification, fatigue resistance, and susceptibility to muscular dystrophy. *Proceedings of the National Academy of Sciences, 112*(2), 424-429. doi:10.1073/pnas.1413021112

Reynolds, S. C., Wilkinson, D. M., Marston, C. G., & O'regan, H. J. (2015, 01). The 'mosaic habitat' concept in human evolution: Past and present. *Transactions of the Royal Society of South Africa, 70*(1), 57-69. doi:10.1080/0035919x.2015.1007490

Rodman, P. S., & Mchenry, H. M. (1980, 01). Bioenergetics and the origin of hominid bipedalism. *American Journal of Physical Anthropology, 52*(1), 103-106. doi:10.1002/ajpa.1330520113

Sawyer, G., & Maley, B. (2005). Neanderthal reconstructed. *The Anatomical Record Part B: The New Anatomist, 283B*(1), 23-31. doi:10.1002/ar.b.20057

Schiaffino, S., & Reggiani, C. (2011, 10). Fiber Types in Mammalian Skeletal Muscles. *Physiological Reviews, 91*(4), 1447-1531. doi:10.1152/physrev.00031.2010

Senut, B., Pickford, M., Gommery, D., & Ségalen, L. (2017, 02). Palaeoenvironments and the origin of hominid bipedalism. *Historical Biology, 30*(1-2), 284-296. doi:10.1080/08912963.2017.1286337

Spangenburg, E. E., & Booth, F. W. (2003, 08). Molecular regulation of individual skeletal muscle fibre types. *Acta Physiologica Scandinavica, 178*(4), 413-424. doi:10.1046/j.1365-201x.2003.01158.x

Talbot, J., & Maves, L. (2016, 05). Skeletal muscle fiber type: Using insights from muscle developmental biology to dissect targets for susceptibility and resistance to muscle disease. *Wiley Interdisciplinary Reviews: Developmental Biology, 5*(4), 518-534. doi:10.1002/wdev.230

Tardieu, C., & Trinkaus, E. (1994, 10). Early ontogeny of the human femoral bicondylar angle. *American Journal of Physical Anthropology, 95*(2), 183-195. doi:10.1002/ajpa.1330950206

Umberger, B. R., Gerritsen, K. G., & Martin, P. E. (2003, 05). A Model of Human Muscle Energy Expenditure. *Computer Methods in Biomechanics and Biomedical Engineering, 6*(2), 99-111. doi:10.1080/1025584031000091678

Wang, Y., Zhang, C., Yu, R. T., Cho, H. K., Nelson, M. C., Bayuga-Ocampo, C. R., . . . Evans, R. M. (2004, 08). Regulation of Muscle Fiber Type and Running Endurance by PPARδ. *PLoS Biology, 2*(10). doi:10.1371/journal.pbio.0020294

Wheeler, P. (1991, 08). The thermoregulatory advantages of hominid bipedalism in open equatorial environments: The contribution of increased convective heat loss and cutaneous evaporative cooling. *Journal of Human Evolution, 21*(2), 107-115. doi:10.1016/0047-2484(91)90002-d

Wheeler, P. (1992, 10). The thermoregulatory advantages of large body size for hominids foraging in savannah environments. *Journal of Human Evolution, 23*(4), 351-362. doi:10.1016/0047-2484(92)90071-g

White, T. D., Suwa, G., & Asfaw, B. (1994, 09). Australopithecus ramidus, a new species of early hominid from Aramis, Ethiopia. *Nature, 371*(6495), 306-312. doi:10.1038/371306a0

Yang, Z. (2007, 04). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Zerbino, Daniel R, et al. (2017). “Ensembl 2018.” *Nucleic Acids Research*, vol. 46, no. D1.

Zipfel, B., Desilva, J. M., Kidd, R. S., Carlson, K. J., Churchill, S. E., & Berger, L. R. (2011, 09). The Foot and Ankle of Australopithecus sediba. *Science, 333*(6048), 1417-1420. doi:10.1126/science.1202703

Zollikofer, C. P., León, M. S., Lieberman, D. E., Guy, F., Pilbeam, D., Likius, A., . . . Brunet, M. (2005, 04). Virtual cranial reconstruction of Sahelanthropus tchadensis. *Nature, 434*(7034), 755-759. doi:10.1038/nature03397